A BOTTOM UP HYBRID METHOD FOR ISOLATED WORDS RECOGNITION

Olivier Delemar & Harouna Kabré
Institut de la Communication Parlée, Grenoble, France

ABSTRACT

In this article we present a new hybrid method for speech recognition which uses expert knowledge to control Hidden Markov Models in a purely bottom-up way. The phonetic knowledge used by the hybrid model to perform this control is embedded in the standard hidden Markov models by the way of an "expert matrix" which expresses whether a given broad phonetic label may or may not, be detected by the expert system while the model is in a given state.

INTRODUCTION

Many recent works in the field of speech recognition tend to combine different methods of acoustic-phonetic decoding, in order to take advantage of their particularities. Among the different hybrid methods, there are those which use the discrimination power of neural networks with the time alignment capacities of Hidden Markov Models [1][2][3] and those which combine rule based systems with HMM [4] or neural networks [5].

The rest of the paper is organized as follows: the first section will present the standard HMM principle, then the hybrid model will be described with an example showing how it controls the recognition algorithm. The last section will give some results and will discuss the enhancements and limitations of this method.

HIDDEN MARKOV MODELING

HMMs are finite states automata which model sequences of quasi-stationary phases [10]. They are formed by a fixed number of states linked to others by arcs. Each arc has an associated transition probability - possibly null - while each state is associated with an emission probability density function (pdf). An N-states Markov model is then defined entirely by its A-matrix of transition probabilities $a_{ij}$ and its set of emission pdf's $b_j(.)$ called the B-matrix.

Speech recognition with HMM consists in evaluating each model probability, given a vector of acoustic features. This may be performed by the Viterbi algorithm [11] which, in addition, finds the best path along the Markov chain in order to maximize this probability. During this process, the choice of a transition from state $i$ to state $j$, given the feature vector $O_\ell$, depends on the value $a_{ij} b_j(O_\ell)$ which may be seen as the "cost" of the transition from state $i$ to state $j$, weighted by the "distance" between the $j^{th}$ state's inner representation of an acoustic configuration and the observed acoustic feature $O_\ell$. Thus the Viterbi algorithm performs nothing other than a time alignment procedure. Figure 1 shows such an alignment for the French word "ouvre" (uvR@).

The hybrid model presented here tries to satisfy three constraints: introducing phonological knowledge expressed by an expert system into the HMM's decision procedure, keeping the automatic aspect of the model's training phase and maintaining the bottom-up aspect of the acoustic-phonetic decoding by HMM which allows a real-time implementation for speech recognition. Taking into account the speech segmentation generated by the Viterbi algorithm, the training phase of the hybrid model will create a so called "expert matrix" $E$ with as many rows as the Markovian model has states and as many columns as there are broad phonetic classes predicted by the expert system.

In our experiments, the expert system is a set of deterministic networks [13] finding occurrences of voiceless fricatives and stop consonants by applying fuzzy thresholds to the zero-crossing rate, the power and its first and second order derivatives. During the standard model training, a time alignment is achieved for each training sample and compared to the broad phonetic labels in order to evaluate the "plausibility" that this label will be predicted by the expert system while the $i^{th}$ state of the model is being visited.

The efficiency of the acoustic-decoding with HMM relies on the optimization of the models' parameters. The forward-backward algorithm [12] which is widely used for this purpose is an Expectation-Maximization procedure which iteratively re-estimates the transition and emission probability, given a training set of acoustic data. It should be noticed that, as far as HMMs are probabilistic models, the number and the quality of the training samples condition the representativity and the generalization capacity of the models. Furthermore, short acoustic events like the stop-consonants' burst are poorly modeled because of their reduced number of representative feature vectors.

The hybrid model has been tested on a subset of the French database BDSON. This corpus is formed by 161 mono or bi-syllabic words, each pronounced once

RESULTS AND DISCUSSION

The hybrid model has been tested on a subset of the French database BDSON. This corpus is formed by 161 mono or bi-syllabic words, each pronounced once
by 5 male and 5 female speakers. The speech signal is sampled at 16kHz and analyzed every 10ms by a Perceptual Linear Predictive Coding algorithm [14] to produce the feature vectors. Both standard and hybrid models are trained with 6 of the 10 utterances and the other 4 are used as the test set.

For this corpus, the standard HMMs give a 38% recognition rate (49% if we consider the first two candidates). By correcting 16% of the confused words, the hybrid model increases this rate to 43% (56% for the first two candidates).

CONCLUSION

We have described a new hybrid approach to speech recognition using HMMs and a rule-based expert system. Phonological knowledge as expressed by the expert system is embedded in the models by the way of the E matrix. This knowledge is then used during a purely bottom-up recognition process, to constrain the states sequence and avoid forbidden states. The results are encouraging but the rules have to be optimized in order to produce solely robust information.

REFERENCES